Book Section:
Dang, TNY orcid.org/0000-0002-3189-7776 (2019) Corpus-Based Word Lists in Second Language Vocabulary Research, Learning, and Teaching. In: Webb, S, (ed.) The Routledge Handbook of Vocabulary Studies. Routledge , pp. 288-303. ISBN 9780429291586

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (ed.), *The Routledge Handbook of Vocabulary Studies(pp.288-304*). New York: Routledge.

Chapter 21. Corpus-based wordlists in second language vocabulary research, learning, and teaching

Thi Ngoc Yen Dang

University of Leeds

ABSTRACT

This chapter examines current research on corpus-based wordlists for second language learners of English. It begins with a review of different kinds of corpus-based wordlists. After that, it discusses four critical issues drawn from the review: the unit of counting, corpus construction, selection criteria, and the reconceptualisation of different kinds of vocabulary. The chapter then discusses how corpus-based wordlists can be applied in L2 learning and teaching. Finally, the chapter concludes by pointing out areas of wordlist studies that deserve further attention, including mid-frequency vocabulary, spoken vocabulary, subject-specific vocabulary, and multi-word units. It also emphasizes the importance of combining objective and subjective criteria in wordlist construction.

INTRODUCTION

Corpora are principled collections of naturally occurring spoken, written, or multimodal data which are stored in electronic format for quantitative and qualitative analysis. Data from large and representative corpora capture actual language use, and therefore, are reliable resources for list developers to identify the words that second language (L2) learners are likely to encounter in

their future use of the target language. Given this benefit, corpora have been widely used to develop and validate wordlists for L2 learners, and these lists have made valuable contributions to multiple aspects of L2 vocabulary learning and teaching. This chapter examines current research on corpus-based wordlists for L2 learners of English. It begins with critical issues and topics in the area and then proposes some directions for future research.

CRITICAL ISSUES AND TOPICS

This section is organised into six sub-sections. The first sub-section provides an overview of different kinds of corpus-based wordlists. The next four sub-sections discuss critical issues drawn from the review: the unit of counting, corpus construction, selection criteria, and the reconceptualisation of different vocabulary types. The last sub-section looks at the application of corpus-based wordlists in L2 vocabulary learning and teaching.

Overview of corpus-based wordlists for L2 learners

Wordlists can be classified into lists of single words and multiwords. Depending on the learning purposes, each type can be further divided into general service lists and specialised wordlists. General service lists focus on the lexical items that are useful for any learning purpose while specialised wordlists draw learners' attention to items that are useful for academic or specific purposes. This section reviews lists of single words and multiwords in turn. Within each type, we will look at general service lists and then specialised wordlists.

Lists of single words

General service words or high-frequency words (e.g., *know, good*) have been widely suggested as the crucial starting point of L2 vocabulary learning. They are small in number, but cover a large proportion of words in different text types. Thus, knowledge of these words may provide

learners with a good chance to comprehend language. Given the importance of these words, a large number of general service lists have been developed. Most of them consist of 2,000-3,000 items. West's (1953) General Service List (GSL)is the oldest and most influential list. It has a long-established status and has had a huge impact on L2 vocabulary learning, teaching, and research. However, developed from texts written in the 1930s, the General Service List may not fully reflect current vocabulary. Subsequent studies have addressed this limitation in two ways. The first way is to develop totally new lists which can potentially replace the GSL: Nation's (2006) BNC2000, Nation's (2012) BNC/COCA2000, Browne's (2013) New General Service List, and Brezina and Gablasova's (2015) New General Service List. The second way is to compile a list with a more manageable size for a specific group of learners from items in existing general service lists: Dang and Webb's (2016a) Essential Word List.

Specialised wordlists can be classified into three types: general academic wordlists, discipline-specific wordlists, and subject-specific wordlists. In this chapter, a discipline refers to a broad group of academic subject areas. For example, academic subject areas such as mathematics and biology can be classified into one disciplinary group (hard sciences) while other subjects such as law and history can be put under another group (soft sciences). The degree of specificity becomes greater from general academic wordlists, discipline-specific wordlists, to subject-specific wordlists.

General academic wordlists are concerned with the shared vocabulary among different academic disciplines. They are useful in English for General Academic Purposes programmes (see Coxhead this volume). To date, Coxhead's (2000) Academic Word List (AWL)is the best-known corpus-based general academic wordlist. This list was developed with the aim to help L2 learners improve comprehension of academic written text, and hashad a great impact on EAP learning

and teaching. However, using West's GSL as the general service vocabulary baseline, the AWLcontains a number of current general service words that the GSL fails to capture (Cobb, 2010; Dang & Webb, 2014). Two new lists have been created with the aim to replace the AWL— Gardner and Davies's (2014) Academic Vocabulary List and Browne, Culligan, and Phillips's (n.d.) New Academic Word List. Two others were also developed to supplement the AWL by focusing on academic spoken discourse— Nesi's (2002) Spoken Academic Word List and Dang, Coxhead, and Webb's (2017) Academic Spoken Word List.

The second type of specialised wordlists—discipline-specific wordlists—narrows their focus on the words that occur across a specific discipline; that is, a group of academic subjects. These lists are suitable for English for Specific Academic Purposesprogrammes. This group includes one written wordlist—Coxhead and Hirsh (2007) EAP Science Word List—and two spoken wordlists—Dang's (2018a) Hard Science Spoken Word List and Dang's (2018b) Soft Science Spoken Word List.

The third type of specialised wordlists—subject-specific wordlists—has the narrowest focus, but has generated a growing interest from researchers. Subject-specific wordlists are also commonly referred to as lists of technical vocabulary (See Liu and Lei this volume). These lists are relevant to English for Specific Purposes (ESP)programmes where all learners plan to study the same subject area (see Chapter 6 for the definition of the term). They focus on the shared vocabulary in a particular subject area, such asagriculture (Martínez, Beck, & Panza, 2009), applied linguistics (Khani & Tazik, 2013; Vongpumivitch, Huang, & Chang, 2009), business (Konstantakis, 2007), chemistry (Valipouri & Nassaji, 2013), medicine (Hsu, 2013; Lei & Liu, 2016; Wang, Liang, & Ge, 2008), engineering (Hsu, 2014; Ward, 1999, Ward, 2009; Watson-

Todd, 2017), nursing (Yang, 2015), or environmental science (Liu & Han, 2015). All of these wordlists have been based on the language found in written texts.

Lists of multiwords

Compared with lists of single words, lists of multiwords are limited in number. Multiwords can refer to any kinds of continuous and discontinuous sequences of words, and how this broad term is interpreted varies from studyto study(see Wood this volume). Most general service lists of multiwords focus on phrasal verbs (Gardner & Davies, 2007; Garnier& Schmitt, 2015; Liu, 2011),collocations (Shin & Nation, 2008), or phrasal expressions (Martinez & Schmitt, 2012).Most specialisedmultiwordlists are general academic wordlists. Three of these general academic wordlistswerederived from written text (Ackerman & Chen, 2013; Byrd & Coxhead, 2010; Durrant, 2009), while the others focused on both spoken and written discourse (Biber, Conrad, & Cortes, 2004; Simpson-Vlach & Ellis, 2010). The focus of these lists also varies from lexical bundles (Biber et al., 2004; Byrd & Coxhead, 2010), formulas (Simpson-Vlach & Ellis, 2010), to collocations (Ackerman & Chen, 2013; Durrant, 2009).

So far this chapter has briefly reviewed different kinds of wordlists, the next sections discuss four critical issues in wordlist studies: the unit of counting, corpus construction, selection criteria, and the reconceptualisation of different vocabulary types.

Unit of counting

One core issue in wordlist studies is what should be counted as a word. Different ways of counting words can influence the results of these studies (Nation, 2016). Words can be seen either as single wordsor multiwords, and the way to count both of these constructsvaries between studies.

Table 1 shows that four common units of counting in lists of single words are:type, lemma, flemma, and level 6 word family. The word type is the number of unrepeated word forms in a text. For example, the sentence, *Counting words is difficult but it is fun*, contains eight wordforms but seven word types because the word form *is*is used twice. The lemma (*admire*) consists of a stem (*admire*) together with its inflected forms (*admires, admired, admiring*). Members of a lemma belong to the same word class. The flemma is similar to the lemma, but it does not distinguish parts of speech. For example, *form* (verb) and *form* (noun) are counted as two lemmas but only one flemma. The word family consists of a stem, together with its potential inflections and derivations that include affixes up to a certain level inBauer and Nation's (1993) scale. This scale is a set of levels (levels 1-7) based on frequency, productivity, predictability, and regularity of affixes. A level 6 word family (*admire*) would includethe stem *(admire*), and members derived from one or more affix up to level 6 (*admirable, admirably, admiration, admirer, admirers, admiringly)* (see Nation, 2016 for more details).

Together with the variation in the unit of counting of single words is the continual debate about the best unit for L2 learners. Nation (2016) points out that all of these units are, in fact, different levels of word families in Bauer and Nation's (1993) scale. Word types shouldbe considered as level 1 word families, lemmas as level 2 families, flemmas as level 2.5, while word families have typically been set at level 6. Therefore, a more accurate question should be which word family level is the most appropriate for a particular group of learners.

The level of word families should match the study purpose (Nation, 2016). Some important questions that need to be answered when determining the unit of counting are (a) who the target users of the list are, (b) whether the list will be used for receptive or productive purposes, and (c) whether the list examines technical words or non-technical words. To find the answer to these

questions, the learning burden should be taken into account. Learning burden is the amount of effort needed to acquire a word family at a certain level (Nation, 2013). The idea behind the concept of word families is that learners may be able to see that word forms with the same stems are related to each other, and therefore, may find it easier to recognise or learn an unknown word. For example, *unhappy*is morphologically related to *happy* rather than a totally unrelated word (*cringe*).

In terms of list users' characteristics, the unit of counting should match the characteristics of learners in a *particular*context. Most existing lists are at either end of Bauer and Nation's (1993) scale. Research has shown that (a) learners from a certain L1 background had difficulty seeing the relationship between stems and derived forms because their L1 does not have these characteristics (Ward, 2009), and (b)learner knowledge of affixes increases with their vocabulary level (Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002). These findings indicate that the most suitable level of counting will vary according to the characteristics of target list users such as their L1 background or their L2 proficiency. Consequently, a certain list may be relevant to a certain group of learners, but may be less applicable to others. In recognition of this need, several recent wordlists (Dang, 2018a, Dang, 2018b; Dang, Coxhead, & Webb 2017; Dang & Webb, 2016a; Gardner & Davies, 2014) chose one word family level as the primary unit of counting but also made the lists available in other units of counting so that the lists would be useful to a wide range of learners.

Table 1. The unit of counting in different wordlists

Kind of wordlists	Unit of counting			
	Word type	Lemma	Flemma	Level 6 word family
General service lists	Essential Word List (Dang & Webb, 2016a)	New-General Service List (Brezina&Gablasova, 2015)	New General Service List (Browne 2013), Essential Word List (Dang & Webb, 2016a)	General Service List (West, 1953), BNC2000 (Nation, 2006), BNC/COCA2000 (Nation, 2012)
General academic word lists	None	Academic Vocabulary List* (Gardner & Davies, 2014)	New Academic Word List (Browne et al., n.d.)	Academic Word List (Coxhead, 2000), Spoken Academic Word List (Nesi, 2002), Academic Spoken Word List** (Dang, Coxhead, & Webb, 2017)
Discipline-specific word lists	None	None	None	EAP Science Word List (Coxhead & Hirsh, 2007), Hard Science Spoken Word List** (Dang, 2018a), Soft Science Spoken Word List** (Dang, 2018b)
Subject-specific word lists	Applied Linguistics Academic Word List (Vongpumivitch, Huang, & Chang2009)	New Medical Academic Word List (Lei & Liu, 2016)	None	Engineering Word List (Ward, 1999), Business Word List (Konstantakis, 2007), Medical Academic Word List (Wang, Liang, & Ge 2008), Agriculture Academic Word List (Martínez, Beck, &Panza 2009), Applied Linguistics Academic Word List (Khani&Tazik, 2013), Medical Word List (Hsu, 2013), Chemistry Academic Word List (Valipouri&Nassaji, 2013), Engineering English Word List (Hsu, 2014), Environmental Academic Word List (Liu & Han, 2015), Nursing Academic List (Yang, 2015), Opaque engineering word list (Watson-Todd, 2017).

*The list is also available in the word family format ** The list is also available in the flemma format

In terms of list purposes, if the list is used for productive purposes, level 1 word families (word types) (Durrant, 2014),level 2 word families (lemmas) (Nation, 2016), or level 2.5 word families (flemmas) are mostsuitable because knowledge of one word form does not mean that learners are able to use derivations of this word productively (Schmitt & Zimmerman, 2002). However, if the study investigates receptive use, word families at higher levels may be more appropriate (Nation, 2016). Learners with knowledge of the base word or one or two members of the word family may be able to recognise other members of the same word-family whenreading or listening. If the study deals with subject-specific technical words, level 1 word families may be most suitable because not all members of a word family are technical words (Chung & Nation, 2004; Nation, 2016).

Let us take Dang and Webb (2016a) Essential Word List (EWL) as an example of how the unit of counting can beselected to match the study purpose. The level 2.5 word family (flemma) was chosen in favour of word families at higher levels for the EWL because it better suits the proficiency level of the target users—L2 beginners, who are still learning the first 1,000 words of English. Because vocabulary size and morphological knowledge are closely related, it is expected that these learners have very limited morphological knowledge, and may not be able to recognise most word family members at higher levels. The level 2.5 word family was chosen over the level 2 word family (lemma) because distinguishing parts of speech may overestimate the learning burden of very closely related items like *smile* (verb) and *smile* (noun). It may not be difficult to infer the meaning of *smile* (noun) in *She has a beautiful smile* if the meaning of *smile* (verb) is known.

Similar to single words, there is variation in the way to count multiwords. As shown in the overview, wordlist studies have examined multiwords from different perspectives: collocations,

lexical bundles, phrasal expressions, phrasal verbs, and formulas. Because Wood (this volume) has discussed this issue in detail, this chapter only emphasisesthatthe variationin the way to interpretmultiwordshas raised the need for a clear definition of this term in each study. Importantly, a framework that systematically draws all units of counting of multiwords together like Bauer and Nation's (1993) word family scale for single words would be useful. Such framework would allow a high degree of consistency in comparing and interpreting results of wordlist studies.

Corpus construction

The nature of corpora has a great impact on the quality of corpus-based wordlists because word selection is primarily based on the information from corpora. This sub-section examines the construction of the source corpora (to develop lists) and validating corpora (to validate lists) in turn.

With advances in technology, new lists have been created from larger corpora with the aim to replace existing lists. This, however, does not necessarily mean that these new lists are better if the corpora from which they weredeveloped are not well-designed and clearly-described. Source corpora should represent as closely as possible the kind of texts that their target users are likely to encounter often in their language use, and detailed descriptions of their composition should be available to users so that they can judge whether the lists developed from these corpora suit their needs. Biber(1993) provides comprehensive guidelines on how to achieve representativeness for corpora in terms of defining the target population, establishing sampling frames and methods, and evaluating the extent to which the final sample represents the full range of text type variation and linguistic variation. Nation (2016) points out that the construction of corpora for wordlist studies should carefully consider the content and the size of corpora, and the nature and size of

the sub-corpora. Unfortunately, while some corpora used to develop wordlists satisfy these guidelines, others do not.

Table 2. Corpora used to develop general service lists of single words

Wordlists	Size (words)	Components	
		Written	Spoken
West's (1953) General Service List	5-million	100%	0%
Nation's (2006) BNC2000	100-million	90%	10%
Nation's (2012) BNC/COCA2000	10-million	40%	60%
Browne's (2013) NGSL	274-million	75.03%	24.97%
Brezina and Gablasova's (2015) New General Service List	12-billion	99.92%	0.08%

Let us take general service lists of single words as an example. These lists aim to capture the words that students are likely to encounter in a range of discourse types. Therefore,an ideal corpus to create a general service list should have equal proportions of spoken and written materials (Nation & Waring, 1997). However, as shown in Table 2, only Nation's (2012) BNC/COCA2000 satisfies this guideline. The BNC/COCA2000 corpus has a good balance between spoken (60%) and written materials (40%) while the corpora used to develop the other lists consist of mainly written materials (75.03%-99.92%). The superiority of the BNC/COCA2000 over the other lists is supported by studies (Dang, 2017; Dang & Webb, 2016b) which compared the five general service lists using lexical coverage, learner vocabulary knowledge, and teacher perceptions of word usefulness as criteria. They found that the BNC/COCA2000 is the most suitable for L2 learners from the perspectives of corpus linguistics, learners, and teachers.

Another example is general academic wordlists. The aim of these lists is to help learners from a wide range of academic disciplines to comprehend academic written/spoken English. Therefore, the corpora to develop these lists should represent materials from various academic disciplines

that these students are likely to encounter frequently in their future study. Importantly, to ensure that the lists are not biased towards the vocabulary in a certain discipline or subject area, the number of subjects per discipline should be the same. So is the number of words per subject. However, as shown in Table 3, except for Coxhead (2000) and Dang et al. (2017), corpora used to develop general academic wordlists were either unclearly described or unbalanced in terms of the number of subjects per discipline and the number of words per subject.

Apart from the corpora used to create wordlists, the corpora used to validate the lists are also important. Testing wordlists in independent corpora provides a valid assessment because the results of the validation are not biased towards the corpora from which the lists were derived (Nation & Webb, 2011). Ideally, the validating corpora should be around the same sizes as those used to develop the lists so that they can capture the vocabulary in the target language as well as the source corpus. Unfortunately, the validating stage is absent from most wordlist studies. Among those having the validation stage, the majority did not meet the mentioned guideline. They either validated wordlists in corpora with smaller sizes than those from which the list was developed or did not test the list in corpora of the same or different genres. For example, of the

Table 3. Corpora used to develop general academic wordlists

Wordlists	Developing corpus						
	Components	Text types	Size	Sub-corpus			
				Number of disciplines	Words/ discipline	Number of subjects/discipline	Words /subject
Coxhead (2000) AWL	100% written	university textbooks, articles, book chapters, laboratory manuals	3.5-million	4	875,000	7	125,000
Browne et al.'s (n.d.) NAWL	98.90% written 1.1% spoken	Journal articles, non-fictions, student essays, lectures, seminars	288-million	unclear	unclear	unclear	unclear
Nesi(2002) SAWL	100% spoken	Lectures, seminars	1.6-million	4	Around 400,000	10-19	2038-36,251
Gardner & Davies's (2014) AVL	100% written	Journal articles, newspapers, magazines	120-million	Not applicable	Not applicable	9	8-22-million
Dang, Coxhead, & Webb's (2017) ASWL	100% spoken	Lectures, seminar, labs, & tutorials	13-million	4	3.5-million	6	500,000

five lists presented in Table 3, only Coxhead's (2000) AWL, Gardner and Davies's (2014) Academic Vocabulary List, and Dang et al.'s (2017) Academic Spoken Word List were tested against independent academic and non-academic corpora. Only the validating corpora of the Academic Spoken Word List had a similar size as the corpus used to develop the list (13-million words). The other validating corpora weremuch smaller than the source corpora— 678,000 words compared to 3.5-million words (Academic Word List) and 16 to 83-million words compared to 120-million words (Academic Vocabulary List).

Selection criteria

The third issue to consider when evaluating corpus-based wordlist studies is the selection criteria. Most studies (e.g., Brezina&Gablasova, 2015; Byrd & Coxhead, 2010) rely solely on objective corpus-driven criteria. Common selection criteria in both lists of single words and multiwords are frequency, range, and dispersion. Frequency is the number of occurrences of a word or a multiword in the entire corpus. Range indicates the number of different texts or sub-corpora that the word/multiword occurs in. Dispersion shows how evenly a word/multiword is distributed across different texts or sub-corpora. The higher frequency, wider range, and more even distribution a word/multiword has, the more likely learners are to encounter it in their language use. Apart from frequency, range, and dispersion, some criteria are unique to the selection of single words ormultiwords. Lexical coverage—the percentage of known words in a corpus—is a common criterion to select single words. The higher the lexical coverage a word provides, the greater value it may have to learners. N-grams, Mutual Information, *t*-scores, and log-likelihood are popular criteria to select multiwords. N-grams refer to the length of the word sequence while the other criteria indicate the extent to which theitems in a certain sequence occur more frequently together than would be expected by chance.

Relying solely on objective corpus-driven criteria to select and validate wordlists has two strengths (Nation, 2016). First, it results in a list that is replicable, which then allows the comparison of the list with other lists using different corpora or different criteria. Second, this approach makes it easier to divide the list into frequency-based sub-lists that can be set as short-term learning goals. However, lists developed from this approach willinevitably be affected by the nature of the corpus from which they were developed. Consequently, some items that may be useful for L2 learners may be absent from these lists because they do not meet the objective selection criteria of the corpus from which the items were derived.

Two approaches have been taken to address this limitation. The first approach is to use objective corpus-driven information as the main criterion in word selection but also use subjective criteria as a guide to adjust these objective criteria so that the lists are more suitable and useful for a particular group of learners. Let us take Dang et al.'s (2017) Academic Spoken Word List as an example. This list is aimed at L2 learners in English for General Academic Purposesprogrammes who plan to study different academic disciplines and have different language proficiency levels. Range, frequency, and dispersion were used as the primary criteria to select items in this list. However, given the limited learning time and slow vocabulary growth rates of L2 learners compared to L1 children, different versions adopting different range, frequency, and dispersion cut-off points were compared. Four key considerations—(a) s*ize and coverage,* (b) *wordfamilies outside general highfrequency word-families,* (c) *distribution across the four disciplinary sub-corpora,* and(d) *adaptability to learners' levels*—were used as the guide to see which version would have the greatest pedagogical value. The first two criteria ensured that the Academic Spoken Word List contained a smaller number of items but provided higher lexical coverage than a general service list. These criteria werenecessary because if the Academic Spoken Word

List either had a larger size but provided less coverage than a general service list, or mainly consisted of general service words, it would not have much value for the target users. It would simply be more useful to learn withexisting general service lists. The last two criteria wereimportant because they made sure that the Academic Spoken Word List couldbenefit the target users irrespective of their academic disciplines and proficiency levels.

Another example is Shin and Nation's (2008) list of high-frequency collocations of spoken English. This list targets elementary learners of English whose vocabulary level is at the most frequency 1,000 words. Frequency of word sequences was used as the main criterion for selection. However, Shin and Nation also made some subjective judgementsby only selecting collocations with node words that werecontent words and among the most frequent 1,000 words. This decision ensured that the selected collocationsweremeaningful units that wouldhave value for teaching and deliberate learning. Thus, by learning items fromthe list, it would help strengthen and enrich knowledge of known words. Moreover, Shin and Nation only selected collocations that met a certain frequency threshold and express a complete meaning so that their list was a manageable size and useful for the users.

The second approach to selecting items is to use subjective criteria together with objective corpus-driven criteria in wordlist development and validation. Let us look at some examples of this approach. West (1953) used one objective corpus-driven criterion (frequency) as the primary criterion to select items for his GSL but also used other subjective criteria (ease of learning, necessity, cover, stylistic level, and emotional neutrality) so that the list wassuitable forL2 learners. Similarly, range, frequency, and dispersion werethe main criteria to select Nation's (2006) BNC2000 and Nation's (2012) BNC/COCA2000 words. However, items that didnot meet the objective criteria but were considereduseful forL2 learnerssuch as informal spoken words

(e.g., *ok, hello*), modern words (e.g., *internet, web*), survival words (e.g., *delicious, excuse*), numbers, weekdays, and months were also included so that these listsweresuitable for L2 learning and teaching. Similarly, Ackerman and Chen (2013)used objective corpus-driven criteria (frequency, range, N-grams, MI, *t*-score, word classes) as the primary criteria to select items for their Academic Collocation List, and then used subjective criterion (experts' judgement) to filter items in the list.

Using corpus-driven information together with subjective criteria may lead to the development of wordlists that better serve L2 learning and teaching purposes compared to solely relying on objective corpus-driven information to develop wordlists. This claim is supported by Dang, Webb, and Coxhead's (under review) study which examined the relationships between lexical coverage, L2 learner knowledge, and English language teacher perceptions of the usefulness of general service words. These researchers found that although these three factors were related,the relationships between lexical coverage and the other two factors were not as strong as that between learner vocabulary knowledge and teacher perceptions. This finding indicates that, although corpora can provide valuable information for wordlist construction, the use of subjective criteria from learners and teachers might be also be necessary to make corpus-based wordlists more relevant to L2 learning and teaching purposes.

Reconceptualisation of different kinds of vocabulary

One issue that arises from the review of wordlist studies is the reconceptualisation of different kinds of words. According to Nation (2001), words can be divided into different 1,000-item bands based on their frequency and range. Items at the first 1,000-word band are the most frequent and wide ranging words while those at the second 1,000-word band have lower frequency and narrower range. The further the 1,000-word bands are from the first 1,000-word

band, the less frequent these bands are. Based on this, Nation (2001) categorised vocabulary into four types—high-frequency words, low-frequency words, academic words, and technical words. High-frequency words are those at the first and second 1,000-word bands. Academic words and technical words are considered as specialised vocabulary and have lower frequency compared to high-frequency words. Low-frequency words are those outside high-frequency words, academic words, and technical words. This traditional classification has been widely accepted, and has had a great influence on wordlist studies. However, recent research has questioned this classification regarding the boundary between high and low-frequency words, and between general and specialised words. Such questions have been reflected in wordlist studies.

Regarding the boundary between high and low-frequency words, 2,000 items has been widely accepted as the number of high-frequency words. As a result, most general service lists (Brezina&Gablasova, 2015; Nation, 2006, Nation, 2012; West, 1953) have around this size. However, investigating the issue from different perspectives (frequency and incidental acquisition, frequency and use of graded readers, and lexicography and dictionary defining vocabulary), Schmitt and Schmitt (2014) suggest that this cut-off point should be expanded to 3,000 (1^{st}-3^{rd} 1,000) words, and items outside general high-frequency words should be divided into mid-frequency words (4^{th}-9^{th} 1,000) and low-frequency words (beyond 9^{th} 1,000). Recognising the limitation of Nation's (2001) classification, Nation (2013) then reconstructed his classification by adding the mid-frequency word level. Compared with the traditional classification, the classification of high, mid, and low-frequency words provides a more precise learning goal for L2 learners at different proficiency stages (seeVilkaite and Schmitt in this volume).

Regarding the boundary between general and specialised vocabulary, following Nation's (2001) classification, the most common approach towards developing specialised wordlists (e.g., Browne et al., n.d.; Coxhead, 2000; Coxhead & Hirsh, 2007; Nesi, 2002)has beento consider academic and technical vocabulary as being outside of general service vocabulary, and therefore, general service words are not included in these lists. For example, Coxhead's (2000) Academic Word List does not include items from West's (1953) list. The strength of this approach is that it considers learners' existing knowledge of general vocabulary, and enables learners and teachers to avoid repeatedly learning and teaching known items. However, specialised wordlists following this approach are inevitably affected by the nature of the general service lists on which they were built.

In recognition of this limitation, a more recent approach has beento consider specialised vocabulary as a separate kind of vocabulary that does not directly relate to general service vocabulary (e.g., Gardner & Davies, 2014; Lei & Liu, 2016; Ward, 1999, Ward, 2009). For instance, Gardner and Davies (2014) included in their Academic Vocabulary List all items that have wider range and higher frequency in academic text than non-academic text that meet the selection criteria even if they are included in existing general service lists. Specialised wordlists following this approach cut across the high, mid, and low-frequency levels,and are not affected by the limitation related to existing general service lists. Yet, as learners' knowledge of general vocabulary is not considered in the list development, there may be a risk of repeatedly teaching and learning known items, which may result in inefficient learning and teaching time.

The third approach (Dang et al., 2017; Dang, 2018a, Dang, 2018b) expands on the two previous approaches. It views specialised vocabulary as a separate kind of vocabulary but also looks at it in the relation to general vocabulary. Let us take Dang et al.'s (2017) Academic Spoken Word

List as an example. Following the secondapproach, general service words were still included in the list if they met the selection criteria. Following the firstapproach, learners' knowledge of general vocabulary was also considered in the list development. However, instead of setting a fixed benchmark of the number of words these learners need to know before learning items from the list, the Academic Spoken Word List wasdivided into four levels according to Nation's (2012) BNC/COCA lists. Depending on their current level of general vocabulary, learners can skip certain levels of the Academic Spoken Word List. This approach proposes a systematic way to integrate specialised vocabulary and general vocabulary. On one hand, it reflects the blurred boundary between general and specialised vocabulary. On the other hand, it ensures that learners' knowledge of general vocabulary is taken into account.

Corpus-based wordlists in L2 learning and teaching

How to apply corpus-based wordlists in L2 curriculum and instruction is a growing concern of current vocabulary research. While there is still a great deal of interest in replacing existing lists with new lists by making use of larger corpora and more powerful statistical measures, an interesting and meaningful move in corpus-based wordlist studies is to make existing lists more suitable to a particular learning and teaching context.

One trend is to create smaller lists thatmatch the proficiency level of a particular group of learners. Ward's (2009) Basic Engineering English wordlist has only 299 items, but may allow its target users (low proficiency foundation engineering undergraduates in Thailand) to recognise around 16% of the words in their textbooks. Dang and Webb's (2016a)Essential Word List consists of the best 800 items (in terms of lexical coverage) from West's GSL, Nation's (2006) BNC2000, Nation's (2012) BNC/COCA2000, andBrezina and Gablasova's (2015) New General Service List. Knowledge of the Essential Word Listheadwords and lemmas may enable itstarget

users (L2 beginners) to recognise 60% and75% of the words in a wide range of spoken and written English texts, respectively.

The second trend is to make a list that is adaptable to learners' proficiency levels (Dang, 2018a, Dang, 2018b; Dang et al., 2017) or break a list into sub-lists with manageable size to fit in individual courses within a language programme(Coxhead, 2000; Dang, 2018a, Dang, 2018b; Dang et al., 2017; Dang & Webb, 2016a). For instance, as discussed previously, Dang et al.'s Academic Spoken Word List is divided into four levels based on lists of general vocabulary, and learners can focus their attention on the Academic Spoken Word List levels that are relevant to their current level of general vocabulary. Eachlevel is further divided into sub-lists with manageable sizes to allow easy incorporation of the list in language learning programmes.

The third trend is to help learners achieve deeper knowledge of items in existing lists. Shin and Nation's (2008) collocation list allows L2 beginners to expand their knowledge of the most frequent 1,000 words from single words to multiwords. Garnier and Schmitt's (2015)PHaVE List supports learners to acquire the most frequent meanings of items in Liu's (2011) phrasal verb list.

Let us now look at the contributions of corpus-based wordlists to different aspects of L2 vocabulary learning and teaching. The most obvious value of these lists can be seen in the area of setting learning goals. Different kinds of words in English do not occur with similar frequencies. Therefore, determining what and how many words need to be learned at different stages of language learning is extremely important. It ensures that learners can get the best return for their learning effort. Wordlists provide course designers with a practical way to decide which words to focus on in a course. The new classification of vocabulary into three levels of high, mid, and low-frequency words provides a more precise long-term learning goals for L2 learner and better

support their continual vocabulary development. Similarly, the reconceptualisation of specialised vocabulary provides EAP/ESP course designers with more flexibility when integrating different kinds of wordlists into the curriculum to meet learners' target subject areas, language proficiency, and learning purposes (see Dang, 2018a for a model of selecting relevant wordlists for different language programme).

The value of corpus-based wordlists has not just been restricted to setting learning goals but also includes the development of vocabulary tests. Tests developed based on corpus-based word lists such as Vocabulary Levels Test (Nation, 1990, Schmitt, Schmitt, &Clapham, 2001; Webb, Sasao, &Ballance, 2017) and Vocabulary Size Test (Nation &Belgar, 2007) have been widely used to assist teachers in setting learning goals and monitoring learners' progress.

With respect to L2 learning and teaching, discussion of wordlists are commonly associated with negative ideas of deliberate rote learning of words out of context. However, it is worth stressing that when effectively done with the right words, such kind of learning can positively contribute to L2 vocabulary development (see Nakata in this volume). Well-designed wordlists, therefore, are essential resources for effective deliberate learning. Importantly, presenting items that learners should learn in the format of wordlists does not mean that these lists should be learned and taught solely through decontextualized methods (Coxhead, 2000; Nation, 2016). Deliberate learning should account for no more than 25% of the amount of time of a well-balanced vocabulary learning programme (Nation, 2007, 2013; Webb & Nation, 2017). The remaining time should be equally spent in learning from meaning-focused input, meaning-focused output, and fluency development. Learning from meaning-focused input means that learners acquire new words through listening or reading activities while learning from meaning-focused output means that learners acquire new words through speaking and writing activities. Fluency development

involves all four skills and does not aim to help students learn new words but to make them become more fluent at using known words. In all three cases, the main focus of attention is on meaning; that is, to understand the messages that they receive or to produce meaningful messages. These strands are only effective if learners have sufficient vocabulary knowledge to make them truly meaning-focused. It means that for learning from meaning-focused input and output to happen, learners should already know at least 95% of the words in the spoken materials (van Zeeland & Schmitt, 2013) and 98% of the words in the written materials (Hu & Nation, 2000). The low proportion of words beyond learners' current vocabulary knowledge allows them to pay more attention to new words and learn them from contexts or with reference to dictionaries or glossaries. Similarly, materials for fluency development activities should involve no or very few unknown words so that learners can process and produce the known words in a fast speed. For these reasons, corpus-based wordlists play a key role in the selection and development of learning materials for meaning-focused input, meaning-focused output, and fluency development activities. Using these lists together with vocabulary processing programmes like RANGE (Heatley, Nation, & Coxhead, 2002), AntwordProfiler (Anthony, n.d), or Lextutor (Cobb, n.d) allows teachers to analyze vocabulary in the materials, and then adapt it so that these materials are relevant to learners' vocabulary levels.

FUTURE DIRECTIONS

There are two directions for future research on corpus-based wordlists. The first direction is to focus more on some under-researched areas of corpus-based wordlists, including mid-frequency vocabulary, spoken vocabulary, subject-specific vocabulary, and multi-word units. The second direction is to bring more pedagogical value to corpus-based research on wordlists. Most corpus-based wordlists rely solely on objective corpus-driven criteria; therefore, future research should

also include other sources of information such as teacher perceptions and learnervocabulary knowledge to support the information from corpora.

FUTHER READING

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. Language Learning, 67(4), 959–997.

This article covers important issues in this chapter: unit of counting, corpus construction, selection criteria, and reconceptualization of general academic vocabulary.

Dang, T. N. Y., & Webb, S. (2016b). Evaluating lists of high-frequency words. *ITL – International Journal of Applied Linguistics*, *167*(2), 132–158.

This article provides comprehensive discussion of high-frequency wordlists and how to evaluate lists of different units of counting.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

This book provides very detailed guidelines for making wordlists for language teaching and testing. It covers important issues in wordlists studies such as identifying the list purposes, deciding the unit of counting, constructing corpora, and criteria to select words.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle, Cengage Learning.

Chapter 8 of this book describes the basic steps of making wordlists.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Chapter 1 of this book explains the definition of corpora, basic steps of making a corpus, and the most common techniques of analysing language in a corpus.

RELATED TOPICS

the different aspects of vocabulary knowledge, defining multi-word items, high, mid, and low frequency words, academic vocabulary, technical vocabulary

REFERENCES

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Anthony, L. (n.d.). *AntwordProfiler*. Retrieved from http://www.laurenceanthony.net/antwordprofiler_index.html

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in University teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36*(1), 1–22.

Browne, C. (2013). The New General Service List: celebrating 60 years of vocabulary learning. *The Language Teacher*, *4*(37), 13–16.

Browne, C., Culligan, B., & Phillips, J. (n.d.). A new academic word list. Retrieved from http://www.newacademicwordlist.org/

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, *5*, 31–64.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*(2), 251–263.

Cobb, T. (n.d.). Lextutor. Retrieved from http://www.lextutor.ca

Cobb, Tom. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language*, *22*(1), 181–200.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé*, *12*(2), 65–78.

Dang, T. N. Y. (2017). *Investigating vocabulary in academic spoken English: Corpora, teachers, and learners* (Unpublished PhD thesis). Victoria University of Wellington, Wellington, New Zealand.

Dang, T. N. Y. (2018a). The hard science spoken word list. *ITL – International Journal of Applied Linguistics*, *169*(1), 44–71.

Dang, T. N. Y. (2018b). The nature of vocabulary in academic speech of hard and soft sciences, *51*, 69–83.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959–997.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.

Dang, T. N. Y., & Webb, S. (2016a). Making an essential word list. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167). Amsterdam: John Benjamins.

Dang, T. N. Y., & Webb, S. (2016b). Evaluating lists of high-frequency words. *ITL – International Journal of Applied Linguistics*, *167*(2), 132–158.

Dang, T. N. Y., Webb, S., & Coxhead, A. (under review). The relationships between lexical coverage, learner knowledge, and teacher perceptions of the usefulness of high-frequency words.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*, 157–169.

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics*, *35*(3), 328–356.

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, *41*(2), 339–359.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645 –666.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range: A program for the analysis of vocabulary in texts*. Retrieved from http://www.vuw.ac.nz/lals/staff/paul-nation/ nation.aspx

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*(4), 454–484.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, *33*, 54–65.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Khani, R., & Tazik, K. (2013). Towards the development of an Academic Word List for applied linguistics research articles. *RELC Journal*, *44*(2), 209–232.

Konstantakis, N. (2007). Creating a business word list for teaching business English. *Elia*, *7*, 79–102.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42–53.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, *45*(4), 661–688.

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, *39*, 1–11.

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183–198.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*(2), 291–304.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language* (1st ed.). Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 1–12.

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle, Cengage Learning.

Nesi, H. (2002). An English Spoken Academic Word List. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the Tenth EURALEX International Congress* (Vol. 1, pp. 351–358). Copenhagen, Denmark. Retrieved from http://www.euralex.org/elx_proceedings/Euralex2002/036_2002_V1_Hilary%20Nesi_An%20English%20Spoken%20Academic%20Wordlist.pdf

Schmitt, N, Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Schmitt, N, & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145–171.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*(01), 17–36.

Shin, D., & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339–348.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, *12*(4), 248–263.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479.

Vongpumivitch, V., Huang, J., & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*(1), 33–41.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*, *27*(4), 442–458.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, *12*(2), 309–323.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170–182.

Watson-Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, *45*, 31–39. https://doi.org/10.1016/j.esp.2016.08.003

Webb, S., & Nation, I. S. P. (2017). *How Vocabulary is Learned*. Oxford: Oxford University Press.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL – International Journal of Applied Linguistics*, *168*(1), 34–70.

West, M. (1953). *A general service list of English words*. London: Longman, Green.

Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*, 27–38.

BIOGRAPHICAL NOTE

Thi Ngoc Yen Dang is a lecturer at the University of Leeds. She obtained her PhD from Victoria University of Wellington. Her research interests include vocabulary studies and corpus linguistics. Her articles have been published in *Language Learning, English for Specific*

Purposes,Journal of English for Academic Purposes, and *ITL-International Journal of Applied Linguistics*.

POSTAL ADDRESS

Thi Ngoc Yen Dang

School of Education

University of Leeds

Hillary Place, Woodhouse Lane

Leeds

LS2 9JT

United Kingdom

Email: T.N.Y.Dang@leeds.ac.uk